ICASE

ON UPSTREAM DIFFERENCING AND GODUNOV-TYPE SCHEMES FOR HYPERBOLIC CONSERVATION LAWS

Amiram Harten

Peter D. Lax

and

Bram van Leer

Report No. 82-5

March 17, 1982

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia 23665

Operated by the

UNIVERSITIES SPACE USRA RESEARCH ASSOCIATION

ON UPSTREAM DIFFERENCING AND GODUNOV-TYPE SCHEMES FOR HYPERBOLIC CONSERVATION LAWS

Amiram Harten*
Tel Aviv University

Peter D. Lax†
Courant Institute of Mathematical Sciences, New York University

Bram van Leer
Institute for Computer Applications in Science and Engineering
and
Leiden State University, the Netherlands

ABSTRACT

This report reviews some of the recent developments in one-sided difference schemes through a unified representation, in order to enable comparison between the various schemes. Special attention is given to the Godunov-type schemes which result from using an approximate solution of the Riemann problem.

Research was supported under NASA Contract No. NAS1-15810 while the first and third authors were in residence at ICASE, NASA Langley research Center, Hampton, VA 23665.

*Funded also by NASA Ames Research Center, Moffett Field, CA under Contract No. NCA2-OR525-001 and the U.S. Department of Energy under Contract No. DE-AC02-76ER03077.

†Funded by the U.S. Department of Energy under Contract No. DE-AC02-76ER03077.

Introduction

Upstream-differencing schemes attempt to discretize hyperbolic partial differential equations by using differences biased in the direction determined by the sign of the characteristic speed.

In recent years upstream differencing has become very popular, and a multitude of new techniques of implementing directionally biased differencing have been suggested. This popularity is primarily due to the robustness of upstream-differencing schemes and the availability of a underlying physical model; other reasons are the possibility of achieving high resolution of stationary discontinuities and that of obtaining fast convergence to a steady state. Most of these schemes are an extension of the Courant-Isaacson-Rees scheme [3] to nonlinear conservation laws, and therefore a unified description may be given.

The present report concentrates on reviewing basic concepts and deriving design principles; numerical experimentation is not presented. The paper is built up as follows: Section 1 reviews some properties of the equations essential to their proper numerical approximation. In Section 2 we discuss a straightforward extension of linear upstream differencing to nonlinear systems. Section 3 introduces the physical picture due to Godunov, useful to interpret certain schemes from Section 2 and to construct new schemes. Finally, in Section 4 we comment on flux splitting, another form of upstream differencing, and relate it to a class of schemes motivated by the Boltzmann equation.

1. Weak Solutions and their Numerical Approximation

In this paper we consider numerical solutions of the initial-value problem for hyperbolic systems of conservation laws

$$u_t + f(u)_x = 0, \quad u(x,0) = u_0(x); \ -\infty < x < \infty. \tag{1.1a}$$

Here $u(x,t)$ is a column vector of m unknowns and $f(u)$, the flux, is a vector-valued function of m components. We can write (1.1a) in matrix form

$$u_t + Au_x = 0, \quad A(u) = f_u \tag{1.1b}$$

(1.1) is called hyperbolic if all eigenvalues of the Jacobian matrix A are real. We assume that the eigenvalues $a_1(u),\ldots,a_m(u)$ are distinct and arranged in an increasing order.

To allow for discontinuous solutions we admit weak solutions which satisfy (1.1) in the sense of distribution theory, i.e.,

$$\int_0^\infty \int_{-\infty}^{\infty} [w_t u + w_x f(u)]\,dxdt + \int_{-\infty}^{\infty} w(x,0)u_0(x)\,dx = 0, \tag{1.2a}$$

for all C^∞ test functions $w(x,t)$ that vanish for $|x| + t$ large.

Condition (1.2a) is equivalent to requiring that for all rectangles $(a,b)\times(t_1,t_2)$ the relation obtained by integrating (1.1a) over the rectangle should hold:

$$\int_a^b u(x,t_2)\,dx - \int_a^b u(x,t_1)\,dx + \int_{t_1}^{t_2} f(u(b,t))\,dt - \int_{t_1}^{t_2} f(u(a,t))\,dt = 0 \tag{1.2b}$$

Clearly, a piecewise smooth weak solution of (1.1) satisfies (1.1) point wise in each smooth region; across each curve of discontinuity the Rankine-Hugoniot relation

(1.3) $$f(u_R) - f(u_L) = S(u_R - u_L)$$

holds, where S is the speed of propagation of the discontinuity, and u_L and u_R are the states on the left and the right, respectively.

Since weak solutions of (1.1) are not uniquely determined by their initial data we select physically relevant solutions, defined as those solutions that are limits as $\varepsilon \to 0$ of solutions $u(\varepsilon)$ of the viscous equations

(1.4) $$u_t + f(u)_x = \varepsilon u_{xx}, \qquad \varepsilon > 0.$$

In this paper we consider systems of conservation laws (1.1) that possess an entropy function $U(u)$, defined as follows:

(i) U is a convex function of u, i.e., $U_{uu} > 0$,

(ii) U satisfies

(1.5a) $$U_u f_u = F_u$$

where F is some other function called entropy flux; it follows from (1.5a) that every smooth solution of (1.1) also satisfies

(1.5b) $$U(u)_t + F(u)_x = 0.$$

Limit solutions of (1.4) satisfy, in the weak sense, the following inequality:

(1.6a) $$U(u)_t + F(u)_x \leq 0;$$

i.e., for all nonnegative smooth test functions $w(x,t)$ of compact support

(1.6b) $$-\int_0^\infty \int_{-\infty}^{\infty} (w_t U + w_x F)\,dx\,dt - \int_{-\infty}^{\infty} w(x,0)\,U(u_0(x))\,dx \leq 0,$$

Condition (1.6b) is equivalent to requiring that for all rectangles $(a,b)\times(t_1,t_2)$ the inequality obtained by integrating (1.6a) over the rectangle should hold:

$$\text{(1.6c)}\qquad \int_a^b U(u(x,t_2))\,dx - \int_a^b U(u(x,t_1))\,dx + \int_{t_1}^{t_2} F(u(b,t))\,dt - \int_{t_1}^{t_2} F(u(a,t))\,dt \le 0.$$

If u is piecewise smooth with discontinuities, then (1.5b) holds pointwise in the smooth regions, while across a discontinuity

$$\text{(1.6d)}\qquad F(u_R) - F(u_L) - S[U(u_R) - U(u_L)] \le 0.$$

Relations (1.6) are called <u>entropy conditions</u> (see [12]).

In the following we shall describe numerical approximations to weak solutions of (1.1) which are obtained by 3-point explicit schemes in conservation form:

$$\text{(1.7a)}\qquad v_j^{n+1} = v_j^n - \lambda f_{j+\frac{1}{2}}^n + \lambda f_{j-\frac{1}{2}}^n \quad , \quad \lambda = \tau/\Delta \ ,$$

where

$$\text{(1.7b)}\qquad f_{j+\frac{1}{2}}^n = f(v_j^n \ , \ v_{j+1}^n).$$

Here $v_j^n = v(j\Delta, n\tau)$, and $f(u,v)$ is a <u>numerical flux</u>. We require the numerical flux to be <u>consistent</u> with the physical flux in the following sense:

$$\text{(1.7c)}\qquad f(u,u) = f(u) \ .$$

We say that the difference scheme (1.7) is consistent with the entropy condition (1.6a) if an inequality of the following kind is satisfied.

$$U_j^{n+1} \leq U_j^n - \lambda F_{j+1/2}^n + \lambda F_{j-1/2}^n \ , \tag{1.8a}$$

where the following abbreviations are used:

$$U_j^{n+1} = U(v_j^{n+1}) \ , \quad U_j^n = U(v_j^n) \tag{1.8b}$$

$$F_{j+1/2}^n = F(v_j^n, v_{j+1}^n) \ ; \tag{1.8c}$$

here $F(u,v)$ is a numerical entropy flux, consistent with entropy flux:

$$F(u,u) = F(u) \ . \tag{1.8d}$$

The following is an easy (but useful) extension of an easy (but useful) theorem of Lax and Wendroff [12]:

Theorem 1.1. Suppose the difference scheme (1.7) is consistent with the conservation law (1.1a), and with the entropy condition (1.6a). Let v_j^n be a solution of (1.7), with initial values $v_j^0 = \phi(j\Delta)$. Extend the lattice function v_j^n to continuous values of x,t by setting, as usual

$$v(x,t) = v_j^n \ , \quad j = [x/\Delta], \quad n = [t/\tau] \ . \tag{1.9}$$

Suppose that for some sequence $\Delta_k \to 0$, $\tau/\Delta = \lambda$, the limit

$$\lim_{\Delta_k \to 0} \quad v(x,t) = u(x,t)$$

exists in the sense of bounded, L_1^{loc} convergence. Then the limit u satisfies the weak form (1.2) of the conservation law, and the weak form (1.6b) of the entropy condition.

The proof consists, just as in [12], of multiplying (1.7a) by a test function, summing by parts over n and j, writing the sum as an integral, and passing to the limit $\Delta_k \to 0$.

Theorem 1.1 remains true, and its proof the same, when the fluxes f and F are allowed to be functions of 2ℓ arguments:

$$f_{j+1/2} = f(u_{j-\ell+1}, u_{j-\ell+2}, \cdots, u_{j+\ell}) \ , \tag{1.10}$$

and similarly for $F_{j+1/2}$.

Assume that $u_0(x)$ is equal to some reference state u_* for $|x|$ large:

$$u_0(x) = u_* \quad \text{for} \quad |x| > M. \tag{1.11a}$$

Then

$$v_j^n = u_* \quad \text{for} \quad \Delta|j| > M + n\Delta \ . \tag{1.11b}$$

The entropy U may be altered by adding to it an arbitrary inhomogeneous linear function; this follows from definition (1.5a). Adding such a linear function to U will not alter

its convexity, but achieves the following:

$$U(u_*) = 0 \ , \quad U_{u_i}(u_*) = 0 \ . \tag{1.12}$$

Since U is convex, it follows from (1.12) that $U(u) > 0$ for $u \neq u_*$; in fact if U is strictly convex,

$$U(u) \geq c|u-u^*|^2 \ . \tag{1.13}$$

Now sum (1.8a) with respect to j over all integers j; we obtain

$$\sum_j U_j^{n+1} \leq \sum_j U_j^n \ . \tag{1.14a}$$

In other words: total entropy is a decreasing function of time. In particular

$$\sum_j U_j^n \leq \sum_j U_j^0 \ . \tag{1.14b}$$

This is an a priori inequality for solutions of the difference scheme (1.7), analogous to the energy inequality for linear symmetric hyperbolic differential and difference equations. Since by (1.13) U is positive for $u \neq u_*$, this is an a priori estimate for solutions of the difference scheme (1.7), and indicates that the scheme is stable. However (1.14b) is not strong enough to prove the pointwise boundedness of solutions of (1.7), nor the existence of convergent subsequences.

A word of caution: when dealing with equations of mathematical physics, in particular the equations of compressible flow, then we must make sure that the difference scheme we are using keeps the variables within their physical range, i.e. that density and pressure are always positive quantities.

Theorem 1.1 holds in any number of space variables. Furthermore multidimensional schemes that are composites of one-dimensional fractional steps satisfy the multidimensional analogue of the entropy condition (1.8a) if each individual one-dimensional step satisfies an entropy inequality of the form (1.8a), see Crandall-Majda, [2].

2. Upstream-differencing Schemes

We start our review with the description of the first order accurate Courant-Isaacson-Rees (CIR) scheme [1] for the constant-coefficient scalar equation

$$u_t + a u_x = 0 , \quad a = \text{const.} \tag{2.1}$$

$$v_j^{n+1} = v_j^n - \lambda a \cdot \begin{cases} v_{j+1}^n - v_j^n & \text{for } a < 0 \\ v_j^n - v_{j-1}^n & \text{for } a > 0 \end{cases} \tag{2.2a}$$

Introducing the notation

$$a^- \equiv \min(a,0) = \frac{1}{2}(a - |a|) ,$$
$$a^+ \equiv \max(a,0) = \frac{1}{2}(a + |a|) ,$$

we rewrite (2.2a) as

$$v_j^{n+1} = v_j^n - [a^+(v_j^n - v_{j-1}^n) + a^-(v_{j+1}^n - v_j^n)] , \tag{2.2b}$$

which can be rewritten as

$$v_j^{n+1} = a^+ v_{j-1}^n + \lambda(1 - |a|)v_j^n - \lambda a^- v_{j+1}^n . \tag{2.2c}$$

Under the Courant-Friedrichs-Lewy (CFL) condition

$$\lambda |a| \leq 1 \tag{2.2d}$$

all coefficients of v_i^n on the right in (2.2d) are positive. Such a scheme is called monotone, and is stable in the maximum norm; that is, for a monotone scheme

$$\max_j |v_j^{n+1}| \leq \max_j |v_j^n| .$$

Equation (2.2d) can be rewritten as

$$(2.2f)\quad v_j^{n+1} = v_j^n - \frac{\lambda}{2} a(v_{j+1}^n - v_{j-1}^n) + \frac{\lambda}{2} |a| (v_{j+1}^n - 2v_j^n + v_{j-1}^n)$$

This shows that solutions of (2.2) can be thought of as approximating solutions of

$$(2.3)\qquad w_t + aw_x = \frac{1}{2} \Delta x |a| (1 - \lambda |a|) w_{xx}$$

to second-order accuracy. We observe that the viscosity term in (2.3) vanishes for $a = 0$; this fact later will allow perfectly resolved stationary shocks but may also result in admitting entropy violating discontinuities.

We describe now the extension of (2.2) to systems of equations with constant coefficients:

$$(2.4)\qquad u_t + Au_x = 0 \ , \quad A = \text{constant}.$$

Because of the hyperbolicity assumption, the system (2.4) can be diagonalized by a similarity transformation

$$(2.5a)\qquad w = T^{-1}u \ , \quad T^{-1}AT = \Lambda, \quad \Lambda_{ij} = a_i \delta_{ij} \ ,$$

$$(2.5b)\qquad w_t + \Lambda w_x = 0 \ .$$

The components of w are called characteristic variables and (2.5b) is a system of decoupled characteristic equations.

We extend the CIR scheme to systems by applying the scalar scheme (2.2) to each of the decoupled scalar characteristic equations. In matrix form this can be written as

$$(2.6a)\quad w_j^{n+1} = w_j^n - \frac{\lambda}{2}\Lambda(w_{j+1}^n - w_{j-1}^n) + \frac{\lambda}{2}|\Lambda|(w_{j+1}^n - 2w_j^n + w_{j-1}^n) ,$$

where the diagonal matrix $|\Lambda|$ is defined by $|\Lambda|_{ij} = |a_i|\delta_{ij}$. In the original variables the scheme (2.6a) takes the form

$$(2.6b)\quad v_j^{n+1} = v_j^n - \frac{\lambda}{2}A(v_{j+1}^n - v_{j-1}^n) + \frac{\lambda}{2}|A|(v_{j+1}^n - 2v_j^n + v_{j-1}^n) ,$$

where $|A| = T|\Lambda|T^{-1}$. Clearly, the stability condition for (2.6a) and (2.6b) is

$$(2.6c)\qquad \frac{\Delta t}{\Delta x}\max_k |a_k| \leq 1 .$$

In general we define the matrix $\chi(A)$ by

$$(2.7)\qquad \chi(A) = T\chi(\Lambda)T^{-1} , \quad (\chi(\Lambda))_{ij} = \chi(a_i)\delta_{ij} .$$

We remark that under our assumption of a full set of eigenvectors we can compute $\chi(A)$ by $\chi(A) = P(A)$, where $P(x)$ is the Lagrangian interpolation polynomial such that $P(a_j) = \chi(a_j)$, $j = 1,\ldots,m$.

In the following we shall describe various techniques to extend the upstream-differencing scheme (2.6b) to nonlinear systems of conservation laws (1.1).

The linear system (2.4) can be regarded as a system of conservation laws (1.1a), where the flux depends linearly on u:

(2.8a) $$f(u) = Au$$

The upwind difference scheme (2.6b) is in conservation form (1.7), with numerical flux given by

(2.8b) $$f(u,v) = A^+u + A^-v$$

where A^+ and A^- are the positive and negative parts of A, defined by the functional calculus (2.7) as

(2.8c) $$A^+ = \chi^+(A) , \qquad A^- = \chi^-(A)$$

where, using (2.2b), we set

(2.8d) $$\chi^+(a) = a^+ , \qquad \chi^-(a) = a^- .$$

Note that since $\chi^+(a) + \chi^-(a) = a$, and $\chi^+(a) - \chi^-(a) = |a|$, we can write

(2.9) $$A^+ = \frac{1}{2}(A + |A|) , \qquad A^- = \frac{1}{2}(A - |A|) .$$

Definition. A difference scheme in conservation form (1.7) is said to be an upstream scheme if:

(i) For u and v nearby states, (2.8b) is a linear approximation to the numerical flux f(u,v).

(ii) When all propagations speeds are > 0,

$$f(u,v) = f(u) ;$$

When all propagation speeds are < 0

$$f(u,v) = f(v) \ .$$

We restate (i) in analytic terms: Suppose u and v are near some reference state u_* ; then we require that

$$(2.10a) \quad f(u,v) = f(u_*) + A^+(u_*)(u-u_*) + A^-(u_*)(v-u_*) + o(|u-u_*| + |v-u_*|) \ .$$

A natural choice for u_* is $\frac{u+v}{2}$; setting this into (2.10a), and making use of (2.9) to set

$$A^+ - A^- = |A|$$

and that

$$f\left(\frac{u+v}{2}\right) = \frac{f(u)+f(v)}{2} + o(|u-v|)$$

we get

$$(2.10b) \quad f(u,v) = \frac{f(u)+f(v)}{2} - \frac{1}{2}\left|A\left(\frac{u+v}{2}\right)\right|(v-u) + o(|u-v|) \ .$$

We can write any numerical flux in the form

$$(2.11a) \quad f(u,v) = \frac{f(u)+f(v)}{2} - \frac{1}{2}\, d(u,v) \ ;$$

for the sake of consistency (1.7c) we need

$$(2.11b) \quad d(u,u) = 0 \ .$$

The upstream condition (2.10b) can be expressed then as

$$(2.11c) \quad d(u,v) = \left|A\left(\frac{u+v}{2}\right)\right|(v-u) + o(|u-v|) \ .$$

Formula (2.6a) shows that linear upstream difference schemes contain a large dose of artificial viscosity, except for those components where a_k is small, in particular where $a_k = 0$. The same appears to be true for all upstream difference schemes for nonlinear conservation laws: when all characteristic speeds are $\neq 0$, each acts like a scheme with a hefty amount of artificial viscosity, smearing discontinuities. There is however quite a distinction among the schemes when one of the characteristic speeds is zero; this shows up in the way each scheme resolves a stationary shock, centered rarefaction wave, and stationary contact discontinuity. We turn now to examining these matters.

The most critical difference in performance occurs in resolving a stationary shock, see (1.6d):

$$(2.12a)\qquad u_0(x) = \begin{cases} u\,, & x < 0 \\ v\,, & x > 0 \end{cases}\,, \qquad f(u) = f(v)\,, \qquad F(v) < F(u)\,.$$

The lack of numerical dissipation allows the design of schemes that perfectly resolve stationary shock, i.e., (2.12a) is a steady solution of the numerical scheme. The condition for that is

$$(2.12b)\qquad d(u,v) = 0 \quad \text{if} \quad f(u) = f(v)\,, \quad \text{and} \quad F(v) < F(u)\,.$$

On the other hand

$$(2.12c)\quad u_0(x) = \begin{cases} u\ , & x < 0 \\ v\ , & x > 0 \end{cases}\ , \qquad F(v) > F(u)$$

is not an admissible discontinuity and should not be a steady solution of the finite difference scheme, i.e., we require that

$$(2.12d)\qquad d(u,v) \neq 0 \quad \text{if} \quad f(u) = f(v)\ , \quad F(v) > F(u)\ .$$

We remark that the danger that a given upstream scheme selects a nonphysical solution will occur only for stationary or near-stationary discontinuities; otherwise there is enough numerical viscosity in (2.3) to enforce the selection of a physically relevant solution. Hence there are two options in designing an upstream differencing scheme for solving problems with discontinuous solution:

(1) To switch direction of differencing in a way that will effectively introduce nonlinear dissipation at the expense of slightly smearing the shock;

(2) To satisfy (2.12) and thus get perfect resolution of stationary shock, but to add a mechanism for checking the admissibility of the discontinuity.

We turn now to describe various forms of $d(u,v)$ in (2.11a). The most straightforward way to generate such functions is by

$$(2.13a)\qquad d(u,v) = |A|(u,v)(v-u)\ ,$$

where $|A|(u,v)$ is a matrix function of u and v which has nonnegative eigenvalues, and such that

(2.13b) $$|A|(u,u) = |A(u)| \ ;$$

$|A(u)|$ is defined by (2.7).

The simplest forms of (2.13) are

(2.14a) $$|A|(u,v) = |A(\tfrac{u+v}{2})| \ ;$$

or

(2.14b) $$|A|(u,v) = \tfrac{1}{2} \ [|A(u)| + |A(v)|] \ .$$

The latter was used by Van Leer in [14], and introduces some nonlinear numerical dissipation that somewhat smears stationary shocks but on the other hand excludes nonphysical discontinuities.

Another form of (2.11c), which has similar properties, has been suggested by Huang, [10]:

(2.15) $$d(u,v) = \mathrm{sgn}\big(A(\tfrac{u+v}{2})\big) \ [f(v) - f(u)] \ ;$$

here $\mathrm{sgn}(x)$ is the sign of x, and $\mathrm{sgn}(A)$ is defined by (2.7).

Yet another type of scheme has been designed by Roe [19]. His scheme is of the form (2.13a), where the matrix function $A(u,v)$ is required to have these properties:

(i)

(2.16a) $$f(v) - f(u) = A(u,v)(v - u)$$

(ii) $A(u,v)$ has real eigenvalues and a complete set of eigenvectors.

(iii)

(2.16b) $$A(u,u) = A(u) \ .$$

For the Euler equations of compressible flow Roe [19] has constructed a linearization of form (2.16a) having these properties. We show now that such a linearization exists quite generally:

Theorem 2.1 (Harten-Lax). Suppose (1.1a) has an entropy function; then (1.1a) has a Roe-type linearization.

Proof: We shall construct an A satisfying (2.16a) which is of form $A = BP$, B symmetric, P positive definite. Clearly, such an A is similar to the symmetric matrix $P^{1/2} S P^{1/2}$ and so has property (ii). In our construction we use the entropy function $U(u)$; since U is convex, the mapping $u \to w = U_u$ is one-to-one; we introduce w as new variable in place of u. Let u_1 and u_2 be two arbitrary states, $w_1 = w(u_1)$, $w_2 = w(u_2)$, $f_1 = f(u_1)$, $f_2 = f(u_2)$. Then

$$(2.17) \quad f_2 - f_1 = \int_0^1 \frac{d}{d\theta} f(\theta w_2 + (1-\theta)w_1)\, d\theta = \int_0^1 f_w \, d\theta\,(w_2 - w_1) = B(w_2 - w_1).$$

We claim that f_w is a symmetric matrix; this implies that S is symmetric. To see about f_w, we use relation (1.5a); differentiating with respect to u we get

$$U_{uu} f_u + U_u f_{uu} = F_{uu} .$$

The second term on the left is a linear combination of symmetric matrices f^i_{uu}; the right side, F_{uu}, also is symmetric. Therefore so is the first term

$$U_{uu} f_u$$

It follows then that also

$$f_u U_{uu}^{-1} \tag{2.18}$$

is symmetric.

Differentiating

$$w = U_u$$

with respect to u shows that $w_u = U_{uu}$; therefore

$$U_{uu}^{-1} = u_w$$

Setting this into (2.18) shows that

$$f_{uu} u_w = f_w$$

is symmetric, as asserted.

Next we express

$$w_2 - w_1 = \int_0^1 \frac{d}{d\theta} w(\theta u_2 + (1-\theta)u_1)\, d\theta$$

(2.19)

$$= \int_0^1 w_u\, d\theta (u_2 - u_1) = P(u_2 - u_1) .$$

Using $w_u = U_{uu}$ and the convexity of U we conclude that P as defined by (2.19) is positive definite.

Combining (2.17) and (2.19) gives

$$f_2 - f_1 = BP(u_2 - u_1) = A(u_2 - u_1) .$$

Thus the A we have constructed can be factored as BP, as asserted. Condition (2.16b) is clearly satisfied.

Note that S depends symmetrically on w_1 and w_2, and P symmetrically on u_1 and u_2. This shows that A is a symmetric function of u_1 and u_2.

A similar result holds for systems of conservation laws in any number of space variables as long as there is an entropy, see Harten [8].

Having constructed A, we can define its absolute value by (2.7); then we set

(2.20) $$d(u,v) = |A(u,v)|(v-u) .$$

When u and v correspond to a stationary discontinuity (2.12), then it follows from $f(v) - f(u) = 0$ and (2.16a) that v-u is a null vector of $A(u,v)$, and consequently in the null space of $|A(u,v)|$; thus $d(u,v) = 0$, whether

or not the entropy condition $F(v) - F(u) < 0$ is satisfied. Therefore the corresponding upstream differencing may admit nonphysical solutions. In an Appendix to [9], Harten and Hyman describe a viscosity-like term that can be added to (2.15) and (2.16) to reject inadmissible discontinuities without affecting the perfect resolution of the physical ones.

Yet another way to construct d in (2.11c) is

$$d(u,v) = \int_u^v |A(w)|\, dw \,, \tag{2.21}$$

where the integration in (2.17) is carried out on a path in state-space connecting u and v. Osher in [16] suggests a path of integration Γ that is piecewise parallel to the right eigenvectors R_k of A :

$$\Gamma = \bigcup_{k=1}^{m} \Gamma_k \,, \tag{2.22a}$$

$$\Gamma_k : \begin{cases} \dfrac{du^k}{d\ell} = R_k(u^k) \,, & , \quad 0 < \ell < \ell_k \,, \\ u^{k+1}(\ell_k) = u^k(0) \,, \end{cases} \tag{2.22b}$$

$$k = 1,\ldots,m.$$

$$u^m(0) = u \,, \quad u^1(\ell_1) = v \,. \tag{2.22c}$$

Existence of a unique solution to (2.22b)-(2.22c) is guaranteed if $\|u-v\|$ is sufficiently small [16]. A consequence

of this choice of path is that d in (2.17) decouples into characteristic contributions

$$d(u,v) = \sum_{k=1}^{m} \int_{0}^{\ell_k} |a_k(u(\ell))| \, R_k(u(\ell)) \, d\ell \, . \tag{2.23}$$

Osher shows that limit solutions of (2.11) with (2.23) satisfy the entropy condition and that a stationary contact discontinuity is perfectly resolved; stationary shocks are smeared over two intermediate states.

3. Godunov-type Schemes

Godunov, in his construction of the "best" monotone scheme [4], has used the exact solutions of local Riemann problems to obtain an upstream difference scheme.

The solution of the Riemann problem

$$(3.1) \qquad u_t + f(u)_x = 0 , \quad u(x,0) = \begin{cases} u_L , & x < 0 \\ u_R , & x > 0 \end{cases} ,$$

depends only on the states u_L and u_R and the ratio x/t ; it will be denoted by $u(x/t; u_L, u_R)$. Since signals propagate with finite velocity,

$$(3.2a) \qquad u(x/t; u_L,u_R) = u_L \quad \text{for} \quad x/t \leq a_L ,$$

$$(3.2b) \qquad u(x/t; u_L,u_R) = u_R \quad \text{for} \quad x/t \geq a_R ;$$

a_L and a_R are the smallest and largest signal velocity.

Godunov derives his scheme by considering the numerical approximation $v(x,t_n)$ of the discrete time levels t_n , $n = 0,1,\ldots,$ to be a piecewise constant function in x, i.e.,

$$(3.3a) \quad v(x,t_n) = v_j^n \quad \text{for} \quad x \quad \text{in} \quad I_j = ((j-\tfrac{1}{2})\Delta, \ (j+\tfrac{1}{2})\Delta) .$$

To calculate the numerical approximation at the next time level $t_{n+1} = t_n + \tau$ we first solve exactly the initial value problem

(3.3b) $\quad u_t + f(u)_x = 0 \ , \quad u(x,t_n) = v(x,t_n) \ , \quad -\infty < x < \infty,$

for $t_n \leq t \leq t_n + \tau$; and denote its solution by $u_n(x,t)$. Each discontinuity in $v(x,t_n)$ constitutes locally a Riemann problem. If we keep $\lambda|a_{max}| < 1/2$ where $|a_{max}|$ is the largest signal speed, then because of (3.2) there is no interaction between neighboring Riemann problems and $u_n(x,t)$ can be expressed exactly in terms of the solutions of local Riemann problems:

$$(3.3c) \quad u_n(x,t) = u\left(\frac{x-(j+\frac{1}{2})\Delta}{t-t_n}; v_j^n, v_{j+1}^n\right) \quad \text{for}$$

$$j\Delta < x < (j+1)\Delta, \quad t_n \leq t \leq t_{n+1} \ .$$

Godunov obtains a piecewise constant approximation $v(x,t_{n+1})$ by averaging $u_n(x,t_{n+1})$, i.e., he sets

$$(3.3d) \qquad v_j^{n+1} = \frac{1}{\Delta} \int_{I_j} u_n(x, t_n+\tau) \, dx \ .$$

We can rewrite (3.3d) in terms of the solutions to the local Riemann problem as

$$(3.4a) \quad v_j^{n+1} = \frac{1}{\Delta} \int_0^{\Delta/2} u(x/\tau; v_{j-1}^n, v_j^n) \, dx + \frac{1}{\Delta} \int_{-\Delta/2}^{0} u(x/\tau; v_j^n, v_{j+1}^n) dx.$$

Since u_n is an exact solution of the conservation laws (3.1) we can evaluate the integral defining v_j^{n+1} in (3.3d) by

applying (1.2b) over $I_j \times (t_n, t_{n+1})$; we get

$$v_j^{n+1} = v_j^n - \lambda [f(\hat{v}_{j+1/2}) - f(\hat{v}_{j-1/2})] , \tag{3.4b}$$

where

$$\hat{v}_{j+1/2} = u(0; v_j^n, v_{j+1}^n) . \tag{3.4c}$$

This shows that (3.4b) is in conservation form, with

$$f(v,w) = f(u(0;v,w)) . \tag{3.4d}$$

The exact solution $u_n(x,t)$ of the Riemann problem satisfies the entropy condition (1.6c):

$$\int_{I_j} U(u_n(x,t_{n+1}))\, dx \leq \Delta U(v_j^n) - \tau F(\hat{v}_{j+1/2}) + \tau F(\hat{v}_{j-1/2}) .$$

Since U is a convex function, Jensen's inequality holds:

$$U\left(\frac{1}{\Delta} \int_{I_j} u(x,t)\, dx\right) \leq \frac{1}{\Delta} \int_{I_j} U(u(x,t))\, dt .$$

Combining the last two inequalities we deduce that Godunov's scheme satisfies the entropy inequality (1.8c).

The description (3.3c) makes sense only if the local Riemann problems don't interact, i.e. if

$$\lambda |a_{max}| < 1/2$$

On the other hand, (3.4b) remains consistent with (3.3d) as long as the waves issuing from $j \pm \frac{\Delta}{2}$ do not react $j \mp \frac{\Delta}{2}$ during the time interval $t_n \leq t \leq t_{n+1}$. This will be the case

as long as

$$\lambda |a_{max}| \leq 1 \ .$$

It follows from the RH relation (1.3) that $f(u(s;v,w)) - su(s;v,w)$ is a continuous function; however it is only piecewise differentiable. It follows that the Godunov flux function $f(v,w)$ defined by (3.4d) is only piecewise differentiable.

Note that Godunov's scheme satisfies criterion (ii) for upstream schemes. To verify that it also satisfies criterion (i) we shall show that Godunov's scheme, when applied to linear equations, reduces to (2.6b).

Consider

(3.5a) $$u_t + A\, u_x = 0 \ ,$$

A a constant matrix. Here the solution of the Riemann problem is composed of constant states separated by a fan of m characteristic lines (see Fig. 1).

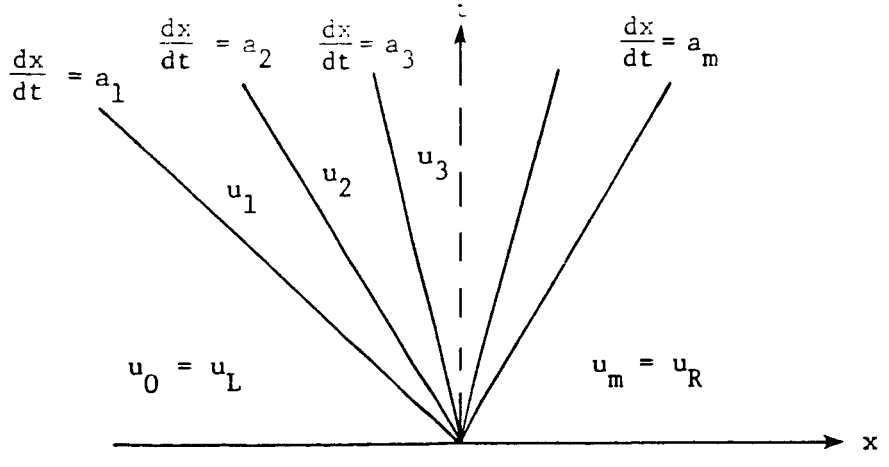

Figure 1.

(3.5b) $u(x/t;u_L,u_R) = u_k$ for $a_k < x/t < a_{k+1}$, $k = 0,\ldots,m$,

where we have defined

$$u_0 = u_L\ ,\ u_m = u_R\ ,\ a_0 = -\infty,\ a_{m+1} = +\infty.$$

The intermediate states u_R can be calculated from the representation of $u_R - u_L$ in terms of the right eigenvectors R_k of A in the following way.

(3.5c) $$u_R - u_L = \sum_{i=1}^{m} J_i R_i\ ,$$

(3.5d) $$u_k = u_L + \sum_{i=1}^{k} J_i R_i\ .$$

We can write this by (2.7) as

(3.5e) $$u_k = u_L + \sigma_k(A)(u_R - u_L)$$

where σ_k is the function

(3.5f) $$\sigma_k(a) = \begin{cases} 1 & \text{for } a < a_k \\ 0 & \text{for } a > a_k \end{cases}$$

Let N be an integer such that

$$a_N < 0 < a_{N+1}$$

Then by (3.5e)

(3.6a) $$u(0;u_L,u_R) = u_N = u_L + \sigma_N(A)(u_R - u_L)$$
$$= (I - \sigma_N(A))u_L + \sigma_N(A)u_R\ .$$

Since by (2.8a), in the linear case $f(u) = Au$, setting (3.6a) into (3.4d) gives

$$f(u,v) = A(I - \sigma_N(A))u + A\sigma_N(A)v$$

From the functional calculus (2.7) and (2.8c) we deduce that

$$A\big(I-\sigma_N(A)\big) = A^+ , \qquad A\,\sigma_N(A) = A^-$$

so that

$$f(u,v) = A^+u + A^-v ,$$

in full agreement with (2.8b). Thus in the linear case Godunov's scheme reduces to the upstream scheme (2.6b).

The solution to the Riemann problem (3.1) has a rather complicated structure: as in the constant coefficient case (3.5) the solution to (3.1) depends on x/t and consists of constant states u_k , $k = 0,\ldots,m$; $u_0 = u_L$, $u_m = u_R$, separated by a fan of waves. Unlike in the constant coefficient case, the k-wave separating u_{k-1} and u_k is not necessarily a single line having a characteristic speed a_k. If the k^{th} characteristic field is genuinely nonlinear then the k-wave is either a rarefaction wave $(a_k(u_{k-1}) < a_k(u_k))$ or a shock propagating with speed S, $(a_k(u_{k-1}) > S > a_k(u_k))$. If the k^{th} characteristic field is linearly degenerate then the k-wave is a contact discontinuity propagating with speed $a_k(u_{k-1}) = a_k(u_k)$ (see [12]).

It is evident from (3.3d) that, due to averaging the Godunov scheme does not make use of all the information contained in the exact solution of the Riemann problem. We therefore consider replacing the exact solution to the Riemann problem $u(x/t;\, u_L,u_R)$ in (3.4a) by an approximation $w(x/t;u_L,u_R)$; the latter can have a much simpler structure as long as it does not violate the essential properties of conservation and entropy inequality. The following theorem due to Harten and Lax [7] (Theorem 2.1) shows that this type of approximation is consistent:

Theorem 3.1 (Harten-Lax). Let $w(x/t;u_L,u_R)$ be an approximation to the solution of the Riemann problem that satisfies the following conditions:

(i) Consistency with the integral form of the conservation law in the sense that

$$\int_{-\Delta/2}^{\Delta/2} w(x/t;u_L,u_R)\, dx = \frac{\Delta}{2}(u_L+u_R) - \tau f_R + \tau f_L \tag{3.7a}$$

for $\Delta/2 > \tau \max |a_k|$, where

$$f_R = f(u_R)\ , \qquad f_L = f(u_L)$$

(ii) Consistency with the integral form of the entropy condition in the sense that

$$\int_{-\Delta/2}^{\Delta/2} U(w(x/t;u_L,u_R))\, dx = \frac{\Delta}{2}(U_L+U_R) - \tau F_R + \tau F_L \tag{3.7b}$$

for $\Delta/2 > \tau \max |a_k|$, where

$$F_R = F(u_R)\ , \qquad F_L = f(u_L)\ .$$

Using the approximation w to the Riemann problem we can define a Godunov type scheme as follows:

$$(3.8)\qquad v_j^{n+1} = \frac{1}{\Delta}\int_0^{\Delta/2} w(x/t; v_{j-1}^n, v_j^n)\,dx + \frac{1}{\Delta}\int_{-\Delta/2}^{0} w(x/t; v_j^n, v_{j+1}^n)\,dx\ .$$

Assertion. If conditions (3.7a) and (3.7b) are satisfied, the scheme (3.8) is in conservation form consistent with (3.1), and satisfies the entropy inequality (1.8a).

For proof, see after Theorem 2.1 in [7]. It is shown there that the Godunov type averaging can be replaced by Glimm type sampling.

Theorem 3.1 shows that Godunov type schemes which satisfy conditions (i) and (ii) above satisfy the hypotheses of Theorem 1.1; this shows that if such a scheme converges, the limit satisfies the conservation law and the entropy condition in the weak sense.

We note that Godunov's scheme is of Godunov type, par excellence.

We have shown at the beginning of this section that Godunov's scheme (3.3) can also be written as a scheme (3.4) in conservation form. The appropriate numerical flux was obtained from the integral conservation laws (1.2b). We show now that all schemes of Godunov type can be expressed

in conservation form; we can obtain the appropriate numerical flux by applying the integral conservation law (1.2b) to the approximate solution of the Riemann problem over the rectangle $(-\frac{\Delta}{2},0) \times (0,\tau)$:

$$(3.9a) \qquad \int_{-\Delta/2}^{0} w(x/\tau;u_L,u_R)\, dx - \frac{\Delta}{2} u_L + \tau[f_{LR} - f_L] = 0 ,$$

where

$$f_{LR} = f(u_L,u_R) .$$

This gives

$$(3.9b) \qquad f_{LR} = f_L - \tau^{-1} \int_{-\Delta/2}^{0} w(x/\tau;u_L,u_R)\, dx + \frac{\Delta}{2\tau} u_L .$$

If we apply the integral conservation law (1.2b) over the rectangle $(0,\Delta/2) \times (0,\tau)$, we obtain

$$(3.9c) \qquad f'_{LR} = f_R + \tau^{-1} \int_{0}^{\Delta/2} w(x/\tau;u_L,u_R)\, dx - \frac{\Delta}{2\tau} u_R$$

The equality of (3.9b) and (3.9c) is just the content of the consistency relation (3.7a).

Using formula (3.9b) for $f_{j+1/2}$ in (1.7b) and (3.9c) for $f_{j-1/2}$ in (1.7b) and setting the resulting expressions into (1.7a) gives (3.8), i.e. puts the Godunov type scheme in conservation form (1.7a):

$$(3.10) \qquad v_j^{n+1} = v_j^n - \frac{\tau}{\Delta} [f(v_j,v_{j+1}) - f(v_{j-1},v_j)] .$$

If all signal speeds are positive, then $w(s;u_L,u_R) = u_L$ for $s < 0$; according to (3.9b), in this case $f_{LR} = f_L$. Similarly, if all signal speeds are negative, then $w(s;u_L,u_R) = u_R$ for $s > 0$; it follows from (3.9c) that in this case $f_{LR} = f_R$. This is property (ii) of upstream schemes, thus shown to be satisfied by all Godunov type schemes.

In a scheme of Godunov type we can incorporate into the numerical flux all physical insight that we can put into the approximate solution of the Riemann problem. Also, as Harten and Lax pointed out in [7], a Godunov type scheme (3.8) can be used just as easily on a grid that varies in time, by adjusting the intervals of integration on the right in (3.8). This makes these schemes the natural choice for adaptive grids; further development of such algorithms and numerical experiments are described in Harten and Hyman [9].

We turn now to describing two different approximate Riemann solvers, and the Godunov type schemes corresponding to them. The first, due to Roe, is based on a linearization motion of type (2.16). Roe approximates solutions of the Riemann problem for (3.1) by exact solutions of the Riemann problem for the following linear hyperbolic equation with constant coefficients:

$$(3.11a) \qquad w_t + A_{LR} w_x = 0, \quad W(x,0) = \begin{cases} u_L & x<0 \\ u_R & x>0 \end{cases}$$

Here $\mathbf{A}_{LR}$ is a matrix that satisfies (2.16a)(2.16b) and has properties i) - iii) listed there. Combining (2.16a) with (3.5c) yields

$$(3.11b) \qquad f_R - f_L = A_{LR}(u_R - u_L) = \sum a_i J_i R_i,$$

where a_i are the eigenvalues of A_{LR}, R_i the corresponding right eigenvectors, and J_i the coefficients in the resolution (3.5c):

$$(3.11c) \qquad U_R - U_L = \sum J_i R_i$$

The approximate Riemann solver is given by (3.5b), with u_k defined by (3.5d).

The numerical flux associated with an approximate Riemann solver is given by (3.9b); setting (3.5d) into (3.9b) we get

$$(3.12a) \qquad f_{LR} = f_L + \sum \bar{a}_i J_i R_i$$

where

$$\bar{a} = \text{Min}\,(a,0) = \tfrac{1}{2}(a - |a|).$$

Setting this into (3.12) and using (3.11b) gives

$$(3.12b) \qquad f_{LR} = \tfrac{1}{2}(f_L + f_R) - \tfrac{1}{2} \sum |a_i| J_i R_i$$

$$= \tfrac{1}{2}(f_L + f_R) - \tfrac{1}{2}|A_{LR}|(u_R - u_L) \ ;$$

in the last step we have used the definition of $|A|$ as given by (2.7). Indeed, (3.12b) is Roe's scheme defined in (2.11a), (2.20).

As already pointed out in Section 2, Roe's scheme admits non-physical, i.e. entropy violating, stationary discontinuities. In an appendix to [9], Harten and Hyman show how to modify Roe's scheme to eliminate such entropy violating discontinuities while retaining those that satisfy the entropy law.

We note that the numerical flux (3.12b) of Roe's scheme resembles Osher's scheme (2.19). There the jumps J_k in the characteristic state variables are represented by the path length ℓ_k. Osher's scheme, however, is not of Godunov type in the sense of (3.9) since the integration path Γ in state-space does not correspond to a univalued approximate Riemann solution $w(x/t,u_L,u_R)$ as in (3.8)

Roe's Riemann solver contains a great amount of detail: m-1 intermediate states. We describe next a hierarchy of Riemann solvers where much of this detail is lumped together. The simples of these schemes contains only one intermediate state.

i) Denote by a_L and a_R lower and upper bounds, respectively, for the smallest and largest signal velocity, calculated according to some algorithm. Define the approximate Riemann solver by

$$(3.13)\qquad u(x/t;u_L,u_R) = \begin{cases} u_L & \text{for } x/t < a_L \\ u_{LR} & \text{for } a_L < x/t < a_R \\ u_R & \text{for } a_R < x/t \end{cases}$$

where the state u_{LR} is determined from the conservation law (3.7a):

$$(\tau a_L+\Delta 12)u_L + \tau(a_R-a_L)u_{LR} + (\Delta 12-\tau a_R)u_R$$

$$= \frac{\Delta}{2}(u_L+u_R) - \tau[f_R-f_L].$$

This gives

$$(3.14) \qquad u_{LR} = \frac{a_R u_R - a_L u_L}{a_R-a_L} - \frac{f_R-f_L}{a_R-a_L}$$

We turn now to the determination of the associated numerical flux. We set (3.14) into (3.13), and then into (3.9b):

$$(3.15a) \quad f_{LR} = \begin{cases} f_L & \text{when } 0 < a_L \\ \dfrac{-a_L}{a_R-a_L} f_R + \dfrac{a_R}{a_R-a_L} f_L + \dfrac{a_L a_R}{a_R-a_L}(u_R-u_L) & \text{when } a_L < 0 < u_R \\ f_R & \text{when } a_R < 0 \end{cases}$$

This can be combined into a single formula

$$(3.15b) \quad f_{LR} = \frac{a_R^- - a_L^-}{a_R-a_L} f_R + \frac{a_R^+ - a_L^+}{a_R-a_L} f_L - \frac{1}{2}\frac{a_R|a_L| - a_L|a_R|}{a_R-a_L}(u_R-u_L)$$

Since u_{LR} was chosen to satisfy the conservation law (3.7a), we conclude that u_{LR} is the mean value of the exact solution over the interval $(\tau a_L, \tau a_R)$. It follows therefore from Jensen's inequality that (3.15) satisfies the entropy inequality (3.7b).

Suppose that u_L and u_R can be connected with a shock of the first or the m^{th} family. In these cases the exact solution is

$$(3.16) \qquad u(x,t) = \begin{cases} u_L \text{ for } x/t < S \\ u_R \text{ for } S < x/t \ , \end{cases}$$

where S is the speed of propagation of the shock. Suppose the algorithm for calculating a_L and a_R is such that it furnishes $a_L = S$ or $a_R = S$, depending on whether the shock belongs to the first or the m^{th} family. Then it follows from the equality of u_{LR} with the mean value of the exact solution that (3.11) is the exact solution.

ii) We describe next a class of approximate Riemann solvers, where u_L and u_R are linked through <u>two</u> intermediate states. These states are so chosen that

a) The conservation laws are satisfied

b) If the exact solution of the Riemann problem links u_L and u_R through a single shock (or contact discontinuity) of any of the m families of waves, then so does the approximate Riemann solver.

c) The entropy law is satisfied

Such an approximate Riemann solver was constructed in [7]; the one presented here differs from it in some important details. We are grateful to Paul Woodward for a suggestion which has been incorporated in the scheme.

Let the velocities a_L and a_R be defined as in approximation i) described above. We define a velocity V as follows: Let U be an entropy function defined by (1.5); denote its gradient by $w=U_u$, and introduce the abbreviation

$$(3.17a) \qquad w(u_R) - w(u_L) = \ell_{LR}$$

We set

$$V = \ell_{LR} \cdot (f_R - f_L) / \ell_{LR} \cdot (u_R - u_L) \tag{3.17b}$$

where the dot denotes the Euclidean scalar product.

Next we show that V is well defined, and derive its salient properties:

Lemma 3.2: i) The denominator in (3.17b) is positive for $u_L \neq u_R$.

ii) V is uniformly bounded.

iii) If u_L and u_R satisfy the RH condition (1.3):

$$f_R - f_L = S(u_R - u_L), \tag{3.18}$$

then V = S.

Proof: i) Combining (3.17a) and (2.19),

$$\ell_{LR} \cdot (u_R - u_L) = P(u_R - u_L) \cdot (u_R - u_L) ; \tag{3.19}$$

since P is positive definite, the above quantity is positive for $u_R \neq u_L$.

ii) Use (2.17), (3.17a) and (2.19) to express the numerator of (3.17b), and (3.19) for the denominator; we get

$$V = \frac{P(u_R - u_L) \cdot BP(u_R - u_L)}{(u_R - u_L) \cdot P(u_R - u_L)}$$

This is a ratio of two quadratic forms and therefore lies between the smallest and largest eigenvalue of $P^{\frac{1}{2}}BP^{\frac{1}{2}}$. These are equal to the eigenvalues a_j of A=BP constructed for Theorem 2.1. In fact, V can be represented as a weighted average of the eigenvalues of A.

iii) Setting (3.18) into (3.17b) gives

$$V = S$$

This completes the proof of Lemma 3.2.

We now outline two methods for constructing approximate Riemann solvers $w(x/t;u_L,u_R)$ with two intermediate states u_L^* and u_R^* separated by the line $dx/dt = V$; i.e., w is of the form

$$(3.20)\quad w(x/t;\, u_L,u_R) = \begin{cases} u_L & x/t < a_L \\ u_L^* & a_L < x/t < V \\ u_R^* & V < x/t < a_R \\ u_R & a_R < x/t \end{cases}$$

(See Figure 2).

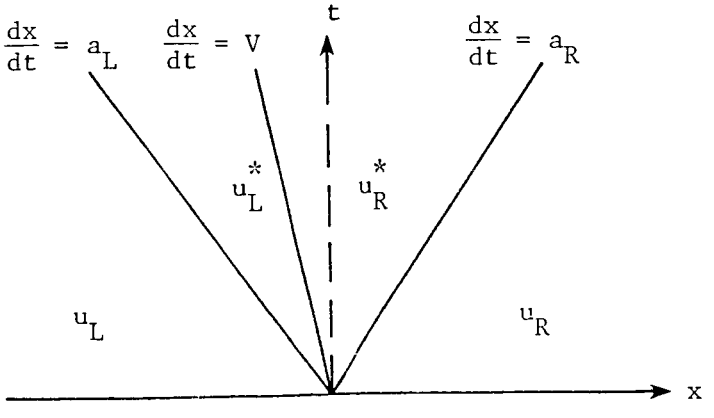

Figure 2.

The flux across a line x=st for equation (1.1) is defined as

$$(3.21)\qquad f_s(u) = f(u)-su$$

We introduce a numerical flux across the line x=Vt, denoted as

$$f_V(u_L,u_R).$$

We require consistency with the exact flux:

(3.22) $$f_V(u,u) = f_V(u) = f(u)-Vu.$$

Having introduced flux (3.21) across lines we can write approximate conservation laws for the triangular regions bounded by $t=\tau$, $x=Vt$, and $x=a_L t$ or $x=a_R t$ respectively:

(3.23a) $$(V-a_L)u_L^* + f_V(u_L, u_R) - f_{a_L}(u_L) = 0$$

and

(3.23b) $$(a_R-V)u_R^* + f_{a_R}(u_R) - f_V(u_L,u_R) = 0$$

u_L^* and u_R^* can be determined from (3.23). Clearly, since (3.23) are conservation laws, the resulting scheme (3.20) satisfied the consistency relation (3.7a). Thus requirement a) is fulfilled.

We turn now to requirement b), the exact resolution of single shocks and contact discontinuities. A shock or contact discontinuity is characterized by the RH condition (3.18) and the entropy condition (1.6d). Using the notation (3.21) these can be written as follows:

(3.24) $$f_S(u) = f_S(u_R), \quad F_S(u_L) \geq F_S(u_R)$$

where

(3.25) $$F_S(u) = F(u)-SV$$

We denote by $\hat{f}_V$ a vector function which has the following property:

If $f_V(u_L) = f_V(u_R)$, then

$$\hat{f}_V(u_L,u_R) = f_V(u_L) = f_V(u_R) \tag{3.26}$$

Below we shall specify two distinct ways for constructing such an $\hat{f}_V$. We now set

$$f_V(u_L,u_R) = \hat{f}_V(u_L,u_R) - \beta(u_L,u_R)(u_R-u_L) \ , \tag{3.27}$$

where β has the following property:

$$\beta(u_L,u_R) = 0 \quad \text{when (3.24) holds.} \tag{3.28}$$

We take (3.27) as our numerical flux. It follows from (3.26) that it satisfies the consistency condition (3.22), and from (3.26), (3.28) and Part c) of Lemma 3.2 that single shocks and contact discontinuities are resolved exactly.

Here are our choices for $\hat{f}_V$ and β: We define δ_R and δ_L by

$$\delta_R = \frac{a_R-V}{a_R-a_L} \ , \qquad \delta_L = \frac{V-a_L}{a_R-a_L} \tag{3.29a}$$

Note that

$$0 \leq \delta_R \ , \qquad 0 \leq \delta_L \ , \qquad \delta_R + \delta_L = 1 \ . \tag{3.29b}$$

Then we define

$$\hat{f}^a_V(u_L,u_R) = \delta_R f_V(u_L) + \delta_L f_V(u_R) \tag{3.30a}$$

and

(3.30b) $$\hat{f}_V^b(u_L,u_R) = f_V(u_{LR}) - f(\delta_L u_L+\delta_R u_R) + \delta_L f_L + \delta_R f_R ,$$

where u_{LR} is defined by (3.14).

It can be verified immediately that f_V^a satisfies (3.26). To verify it for f_V^b we note that if $f_V(u_L)=f_V(u_R)$, then $u_{LR} = \delta_R u_R+\delta_L u_l$; setting this into the right side of (3.30b) we see that it equals the right side of (3.30a).

We define β as follows:

(3.31a) $$\beta = C_1\beta_1+C_2\beta_2$$

where

(3.31b) $$\beta_1(u_L,u_R) = \left[F_V(u_R) - F_V(u_L) - \frac{1}{2}\,(u_L+u_R)\cdot(f_V(u_R)-f_V(u_L))\right]^+ \|u_R-u_L\|^{-2}$$

where p^+ denote Max(0,p), and

(3.31c) $$\beta_2(u_L,u_R) = (a_R-a_L)^{-1}\,\|f_V(u_R)-f_V(u_L)\|^2\,\|u_R-u_L\|^{-2}$$

The analysis in Section 4 of [7] shows β_1 is a bounded function.

By construction, $\beta_1 = 0$ when the shock condition (3.24) is satisfied; $\beta_2 = 0$ when the RH condition (3.18) alone is satisfied. Thus our choice of β satisfies (3.28), and so requirement b) is fulfilled.

An analysis similar to that carried out in Section 4 of [7] shows that the positive constant C_1 and C_2 in (3.31a) can be so chosen that the entropy condition (3.7b) is satisfied. Thus is requirement c) fulfilled.

To derive the numerical flux associated with the above Godunov type scheme we set (3.20) in (3.9b), using (3.23) to express u_L^*

and u_R^*:

$$(3.32a)\qquad f_{LR} = \tfrac{1}{2}\{ f_L+f_R+\gamma_L[f_V(u_L)-f_V(u_L,u_R)] + \gamma_R[f_V(u_L,u_R)-f_V(u_R)]-|V|(u_R-u_L)\}$$

where

$$(3.32b)\qquad \gamma_L = \frac{|V|-|a_L|}{V-a_L}\ ,\ \gamma_R = \frac{|a_R|-|V|}{a_R-V}$$

One can easily verify that when $a_L \geq 0$, $f_{LR} = f_R$, and that when $V = 0$, $f_{LR} = f_V(u_L,u_R)$.

If on the right side of (3.32a) we substitute (3.27) for $f_V(u_L,u_R)$, the following term containing β appears:

$$(3.33a)\qquad -\tfrac{1}{2}(\gamma_R-\gamma_L)\beta(u_R-u_L)$$

From (3.32b)

$$(3.32b)\qquad \tfrac{1}{2}(\gamma_R-\gamma_L) = \begin{cases} 0 & \text{if} \quad a_L > 0 \text{ or } a_R < 0 \\ \dfrac{|a_L|}{|a_L|+|V|} & a_L < 0 < V < a_R \\ \dfrac{|a_R|}{|a_R|+|V|} & a_L < V < 0 < a_R \end{cases},$$

a nonnegative quantity. This shows that β enters the difference scheme as an artificial viscosity. Note that, unlike classical artificial viscosity, our β is zero across a shock and is positive across an incipient rarefaction wave.

Unlike the previous schemes described in this review, the schemes (3.32) are nonlinear even when applied to linear equations. Thus it is not upstream in the sense of our definition in Section 2. The decrease of entropy guarantees the L_1 stability of the scheme.

For systems with many components, (3.32) requires less computational effort than either Godunov or Roe's scheme, yet its accuracy may be comparable.

Schemes of type (3.32) are especially suitable for computation on a moving mesh: we move each meshpoint with velocity V. Such mesh algorithms have been studied in [9].

We remark that any scheme in conservation form (1.7) with a numerical flux f(u,v) that yields perfect resolution of discontinuities but also admits entropy violating ones may be corrected by modifying its numerical flux to be

$$f(u,v) - C_1\beta(v-u) .$$

4. Flux-Splitting

In this section we discuss generalizations of the upstream-differencing scheme (2.6b) to nonlinear systems of conservation laws that are based on the flux-splitting

$$f(w) = f^{+}(w) + f^{-}(w) \ . \tag{4.1a}$$

We consider schemes in conservation form (1.7a) with the numerical flux

$$f(u,v) = f^{+}(u) + f^{-}(v) \ ; \tag{4.1b}$$

clearly (4.1a) implies the consistency relation (1.7c). Let us define

$$f^{a}(w) = f^{+}(w) - f^{-}(w), \tag{4.2a}$$

and rewrite (4.1b)

$$f(u,v) = \frac{1}{2}\left[f(u)+f(v)-\left(f^{a}(v) - f^{a}(u)\right)\right]. \tag{4.2b}$$

Recalling the notation (2.11a) we write

$$d(u,v) = f^{a}(v) - f^{a}(u) \ . \tag{4.2c}$$

It is easy to see that (4.1)-(4.2) reduces to (2.6b) in the constant-coefficient case if and only if $f^{a}(w)$ becomes $|A|w$.

Steger and Warming [21] introduced the notion of flux splitting for the equations of gas dynamics. They took advantage of the fact in gas dynamics f is a homogeneous function of w of degree one. Then Euler's identity holds:

$$f(w) = A(w)w\ , \qquad A = f_w. \tag{4.3a}$$

Steger and Warming define

$$f^+(w) = A^+(w)w\ , \qquad f^-(w) = A^-(w)w \tag{4.3b}$$

where A^+ and A^- are defined by (2.7), (2.9). Clearly (4.1a) is satisfied; (4.2a) becomes

$$f^a(w) = |A(w)|w\ , \tag{4.3c}$$

where $|A(w)|$ is defined by (2.7).

Setting (4.3c) into (4.2c) gives, after rearrangement

$$d(u,v) = \frac{1}{2}\left(|A(u)| + |A(v)|\right)(v-u) + \frac{1}{2}\left(|A(v)| - |A(u)|\right)(u+v). \tag{4.3c'}$$

The RHS is of the upstream form (2.11c'), except for those nearby values of u and v for which sgn $a_k(u) \neq$ sgn $a_k(v)$ for some k. The consequence of this nonsmoothness is a kink in computed solutions near such transitions, e.g. near sonic points. This can be rectified to

some extent if one replaces $|A|$ by $\chi(A)$, where $\chi(s)$ is a smooth approximation to $|s|$, see Steger [22], Van Leer [15], or Harten [5].

We describe now a way of splitting when f(u) is not a homogeneous function of u, by introducing a reference state u_0 and applying the mean value linearization (2.16a) to u and u_0:

$$f(u) - f(u_0) = A(u_0,u)(u - u_0) \tag{4.4a}$$

and consider $u-u_0$ to be a new state variable, and $f(u)-f(u_0)$ to be a new flux function; denote them again by u and f(u), respectively. Thus any solution of the conservation law (1.1) satisfies

$$u_t + [A(u_0,u)u]_x = 0 \tag{4.4b}$$

The new flux

$$f(u) = A(u_0,u)u \tag{4.5a}$$

can be split as

$$f(u) = A^+(u_0,u)u + A^-(u_0\, u)u = f^+(u) + f^-(u),$$

exactly as in the homogeneous case (4.3a).

We turn now to a class of upstream schemes which are a natural generalization of Steger and Warming's flux splitting. These schemes

are obtained by approximating each conservation law in (1.1a) by a collisionless Boltzmann equation.

Let $\phi(x,t,q)$ be a vector function whose i^{th} component denotes the density of a particle of i^{th} kind at position x and time t, travelling with velocity q. We assume that the particles stream freely, i.e. that ϕ satisfies the collisionless Boltzmann equation:

(4.6a) $$\phi_t + q\,\phi_x = 0 .$$

Note that the densities of the different kinds of particles are completely decoupled.

Using equation (4.6a) we can determine the value of ϕ for $t > 0$ in terms of $\phi_0(x,q) = \phi(x,0,q)$:

(4.6b) $$\phi(x,t,q) = \phi_0(x-qt,q).$$

We denote by z and g the total density and flux associated with the density ϕ:

(4.7a) $$z(x,t) = \int \phi(x,t,q)\, dq$$

(4.7b) $$g(x,t) = \int q\, \phi(x,t,q)\, dq$$

These satisfy the conservation law obtained by integrating (4.6a) with respect to q:

(4.8) $$z_t + g_x = 0$$

The initial values of z and g can be obtained by setting t = 0 in (4.7):

(4.9) $$z_0(x) = \int \phi_0(x,q)\, dq\ , \quad g_0(x) = \int q\ \phi_0(x,q)\, dq$$

If the initial values of z and g are equal to those of u and f(u):

(4.10) $$z_0(x) = u_0(x)\ , \qquad g_0(x) = f(u_0(x)),$$

then for t small enough the solution z of (4.8) would be a reasonable approximation to the solution u of (1.1a). We show now how to choose ϕ_0 so that (4.10) holds: we introduce a vector distribution $\mu(q,u)$, depending on a vector parameter u, satisfying

(4.11a) $$\int \mu(q,u)\, dq = u\ ,$$

(4.11b) $$\int q\ \mu(q,u)\, dq = f(u)\ .$$

Then we simply set

(4.12a) $$\phi_0(x,q) = \mu(q,u_0(x)).$$

Setting (4.12a) into (4.9) and using (4.11) shows that (4.10) is satisfied.

Equation (4.8) can be solved explicitly; as in the Godunov type schemes, averages of these explicit solutions will be used to approximate solutions of (1.1a).

Let $v(x)$ be an approximation at t_n to the solution u of (1.1a). We define the initial value ϕ_0 by (4.12a):

$$\phi_0(x,q) = \mu(q, v(x)) \,. \tag{4.12b}$$

Using this in formula (4.6b) for the exact solution of equation (4.6a) gives

$$\phi(x,t,q) = \mu(q, v(x-qt)) \,. \tag{4.12c}$$

Setting this into (4.7a) results in

$$z(x,t) = \int \mu(q,v(x-qt))\, dq. \tag{4.13}$$

We assume that $v = v^n$ is piecewise constant. We define an approximation v^{n+1} to u at $t_{n+1} = t_n+\tau$ that is piecewise constant on each interval $I_j = (x_{j-1/2}, x_{j+1/2})$ by defining the value v_j^{n+1} of v^{n+1} on I_j to be the average of $z(x,\tau)$ over I_j. Using (4.13) we get

$$v_j^{n+1} = |I_j|^{-1} \iint_{I_j} \mu(q,v^n(x-q\tau))\, dq\, dx \tag{4.14}$$

We show next how to express (4.14) as a scheme in conservation form. We integrate (4.6a) over the rectangle $I_j \times (0,\tau)$; we get

$$\int_{I_j} \phi \, dx \Big|_0^\tau + \int_0^\tau q \, \phi \, dt \Big|_{x_j-1/2}^{x_{j+1/2}} = 0$$

We integrate this with respect to q, obtaining the conservation form

$$(4.15a) \quad |I_j|(v_j^{n+1} - v_j^n) + \tau(f_{j+1/2}^n - f_{j-1/2}^n) = 0 \,,$$

where v_j^{n+1}, v_j^n are the values of v^{n+1}, v^n on I_j, and $f_{j\pm 1/2}^n$ is defined by

$$\tau \, f_{j+1/2}^n = \iint_0^\tau q \, \phi(x_{j+1/2}, t, q) \, dt \, dq$$

Using (4.12c) to express ϕ on the RHS we get

$$(4.15b) \quad \tau \, f_{j+1/2}^n = \int_0^\tau \int q \, \mu(q, v^n(x_{j+1/2} - qt)) \, dt \, dq.$$

This is the numerical flux associated with the scheme (4.14).

We show now that this numerical flux is consistent with the flux f in (1.1a). Let's take the case, sure to be satisfied in any scheme of practical significance, that μ has bounded q-support. Then it follows from (4.15b), and the fact that v^n is piecewise constant, that there is an integer N, whose value depends on τ, such that

$$f_{j+1/2}^n = f(v_{j-N+1}^n, \ldots, v_{j+N}^n)$$

Suppose $v^n_{j+k} = v$, $k = -N+1,\ldots,N$; then $v^n(x_{j+1/2}-qt)$ on the RHS of (4.15c) equals v on the q-support of μ; using (4.11b) we see that the RHS of (4.15b) equals $\tau f(v)$. This proves consistency.

We turn now to the task of determining the distribution μ. Clearly, since only the first two moments of μ are specified by conditions (4.11), there is a great deal of leeway. For guidance we turn to the linear case,

$$f(u) = Au, \quad A \text{ constant}$$

The solution of the linear equation

$$u_t + Au_x = 0 \,, \tag{4.16a}$$

with initial value $u(x,0) = u_0(x)$ has the form

$$u(x,t) = \sum P_i u_0(x - a_i t). \tag{4.16b}$$

Here a_j are the eigenvalues of A, and P_i is projection onto the line spanned by the right eigenvector Ri. On the other hand, setting (4.6b) into (4.7a) gives the following expression for z:

$$z(x,t) = \int \phi_0(x-qt,q)\, dq \tag{4.17}$$

Clearly, comparing (4.16b) and (4.17) we see that

$$z(x,t) \equiv u(x,t) \tag{4.18a}$$

if

(4.18b) $$\phi_0(x,q) = \sum P_i\, u_0(x)\, \delta(q - a_i).$$

It is not hard to show that (4.18a) holds for no other choice of ϕ_0 whose support is bounded in x and q. Setting (4.18b) for ϕ_0 into (4.12a) we conclude that μ must be of the form

(4.19) $$\hat{\mu}(q,u) = \sum \delta(q-a_i)P_i u$$

We turn now to the nonlinear case. As we have shown earlier in this section, the flux can be put in form (4.5a):

$$f(u) = A(u)u,$$

where the matrix A(u) has real eigenvalues $a_i(u)$, and a complete set of eigenvectors. According to the spectral theory of matrices

(4.20) $$\sum P_i = I\ , \qquad \sum a_i P_i = A$$

It follows from this that the distribution $\hat{\mu}$ defined by (4.19) satisfies relations (4.11) even when a_i and P_i are functions of u.

Since relations (4.11) are linear, the most general μ that satisfies (4.11) is of the form

(4.21) $$\mu(q,u) = \hat{\mu}(q,u) + \eta(q,u),$$

where η is any distribution whose first two q-moments are zero for all values of u.

We call schemes of form (4.14) Boltzmann-like; we list some of their properties.

(a) Suppose the support of the distribution μ is contained in $|q| \leq Q$. Suppose for simplicity that the mesh over which we discretize is uniform, i.e. that each interval I_j has the same length Δ. Then it follows easily that (4.14) is a three point scheme if τ is chosen so that

$$\tau Q \leq \Delta. \tag{4.22}$$

For Boltzmann-type schemes the flux (4.15b) splits naturally into two parts. We define

$$\mu_+(q,u) = \begin{cases} \mu(q,u) & \text{for } q \geq 0 \\ 0 & \text{for } q < 0 \end{cases}$$

$$\mu_-(q,u) = \begin{cases} 0 & \text{for } q > 0 \\ \mu(q,u) & \text{for } q < 0 \end{cases} \tag{4.23}$$

Clearly $\mu = \mu_+ + \mu_-$; setting this into (4.15b) we obtain a splitting

$$f_{j+1/2} = f^+_{j+1/2} + f^-_{j+1/2}$$

Clearly, if (4.22) holds, $f^+_{j+1/2}$ depends only on v_{j+1}, and $f^-_{j+1/2}$ only on v_j.

(b) Suppose f is of the form (4.5a) and μ is chosen to be of form (4.19). The support of $\hat{\mu}$ extends from a_{min} to a_{max}, so $Q = |a_{max}|$, and the restriction (4.22) is the CFL condition. The decomposition of $\hat{\mu}$ is

$$\hat{\mu}_+ = \sum_{a_i \geq 0} \delta(q - a_i)\, P_i u$$

$$\hat{\mu}_- = \sum_{a_i < 0} \delta(q - a_i)\, P_i u$$

Setting this into (4.15b) gives

$$f^+_{j+1/2} = A^+(v_{j+1})v_{j+1}$$

$$f^-_{j+1/2} = A^-(v_j)v_j$$

where A^+ and A^- are the positive and negative parts of the matrix A defined by (2.7) and (2.9). This is the Steger-Warming scheme (4.3).

(c) Consider the equations of compressible flow in Euler coordinates. In this case the three components of ϕ describe the transport of mass, momentum and energy. Since momentum is mass × velocity, and q is velocity, it is reasonable to stipulate that

the second component of ϕ be q times its first component. In view of (4.12a), this would be the case iff the same relation holds for the components of μ:

$$\mu^{(2)} = q\ \mu^{(1)}.$$

We claim that this is true for $\hat{\mu}$ as given by (4.19). For

$$P_i u = w_i R_i$$

and so, by (4.19),

(4.23a) $$\mu = \sum \delta(q - a_i) w_i R_i$$

(4.23b) $$q\hat{\mu} = \sum \delta(q - a_i) a_i w_i R_i$$

For the equations of compressible flow, the first component $f^{(1)}$ of flux and the second conserved quantity $u^{(2)}$ both are equal to m. This implies that the first row of $A = f_u$ is (0,1,0). It follows then from the eigenvalue equation

$$AR_i = a_i R_i$$

that the second component of R_i is a_i times its first component. This shows that the second component of (4.23a) equals the first component of (4.23b), as asserted.

(d) We analyze now the stability of Boltzmann type schemes for linear equations with constant coefficients. In this case we take μ to depend linearly on u,

$$\mu(q,u) = M(q)u, \tag{4.24}$$

M a matrix valued function. The consistency conditions (4.11) become

$$\int M(q)\,dq = I\ , \qquad \int q\,M(q)\,dq = A. \tag{4.25}$$

The scheme (4.13) is

$$z(x,t) = \int M(q)\,v(x - qt)\,dq$$

Taking the Fourier transform we get

$$\hat{z}(\xi,t) = \hat{M}(t\xi)\hat{v}(\xi)$$

A condition for stability is that $\hat{M}(\xi)$ should be power bounded, uniformly for all ξ. Note that there is no stability restriction of the time step t; the reason for this is that Boltzmann type schemes automatically adjust their domains of dependence.

Note that we have analyzed here the stability of scheme (4.13). The full scheme (4.14) is a combination of (4.13) and projection onto the space of piecewise constant functions. The latter decreases every weighted L_2 norm. Therefore if (4.13) decreases some weighted L_2 norm, the combined scheme (4.14) is L_2 stable.

For the case (4.19), we have

$$M(q) = \sum \delta(q-a_i)P_i,$$

and

$$\hat{M}(\xi) = \sum e^{iq_j\xi} P_j$$

If A is symmetric, the P_j are orthogonal projections, and $\|\hat{M}(\xi)\| \equiv 1$; so in this case the scheme is stable. For A nonsymmetric, stability can be proved by replacing the euclidean norm by some matrix-weighted norm.

(e) We have not carried out any stability analysis in the nonlinear case, nor studied the interesting question of how to assure the entropy condition.

(f) We conclude by observing that flux splitting schemes cannot resolve exactly stationary shocks. For suppose that the stationary RH condition $f(u) = f(v)$ is satisfied. It does not follow from this that also

$$f^a(u) = f^a(v). \tag{4.26}$$

But for a split flux scheme, (4.2b) shows that (4.26) is necessary for the exact resolution of stationary discontinuities.

References

[1] R. Courant, E. Isaacson, and M. Rees, *Comm. Pure Appl. Math.*, 5 (1952), pg. 243.

[2] M.G. Crandall, A. Majda, *Math. Comp.*, 34, (1980), pp. 1-21.

[3] J. Glimm, *Comm. Pure Appl. Math.*, 18 (1965), pp. 697-715.

[4] S.K. Godunov, *Math Sbornik*, 47 (1959), pg. 271; also: Cornell Aero. Lab. Transl.

[5] A. Harten, "High Resolution Schemes for Hyperbolic Conservation Laws", New York Univ. Report, DOE/ER/03077-167, March 1982.

[6] A. Harten, J.M. Hyman, and P.D. Lax, *Comm. Pure Appl. Math.*, 29 (1976), pp. 297-322.

[7] A. Harten and P.D. Lax, "A Random Choice Finite-Difference Scheme for Hyperbolic Conservation Laws", *SIAM J. Num. Anal.*, v. 18, 1981, pp. 289-315.

[8] A. Harten, "On the Symmetric Form of Systems of Conservation Laws with Entropy", ICASE Report No. 81-34, 1981.

[9] A. Harten, J.M. Hyman, "A Self-Adjusting Grid for the Computation of Weak Solutions of Hyperbolic Conservation Laws", Center for Nonlinear Studies Theoretical Division, Los Alamos Nat. Lab. LA9105, (1981).

[10] L.C. Huang, "Pseudo-Unsteady Difference Schemes for Discontinuous Solutions of Steady-State One-Dimensional Fluid Dynamics Problems", *J. Comp. Phys.*, 42 (1981) 195-211.

[11] S. Kaniel and J. Falcovitz, "Transport Approach for Compressible Flow", IRIA Conference on Numerical Methods, Paris, 1979, to appear in Springer-Verlag Lecture Notes.

[12] P.D. Lax, "Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves", *SIAM*, Philadelphia, 1972.

[13] B. van Leer, *J. Comp. Phys.*, 23, (1977), pp. 263-275.

[14] B. van Leer, "Upwind Differencing for Hyperbolic Systems of Conservation Laws", Invited Lecture, 2nd Int'l. Congress on Numerical Methods in Engineering, Paris, December 1-5, 1980.

[15] B. van Leer, "Flux-Vector Splitting for the Euler Equations", ICASE Report, to appear.

[16] S. Osher, "Numerical Solution of Singular Perturbation Problems and Hyperbolic Systems of Conservation Laws", UCLA preprint 1980, to appear.

[17] D.J. Pullin, *J. Comp. Phys.*, 34 (1980), pp. 231-244.

[18] R.D. Rietz, "One-dimensional Compressible Gas Dynamics Calculations Using the Boltzmann Equation", August 1980, to appear in *J. Comp. Phys.*

[19] P.L. Roe, Proc. Seventh International Conference on Numerical Methods in Fluid Dynamics, Stanford/NASA Ames, June 1980, Springer-Verlag.

[20] P.L. Roe, "Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes", *J. Comp. Phys.*, 43 (1981), pp. 357-372.

[21] J.L. Steger and R.F. Warming, "Flux Vector Splitting of the Inviscid Gasdynamic Equations with Applications to Finite Difference Methods", July 1979, NASA TM-78605.

[22] J.L. Steger, "A Preliminary Study of Relaxation Methods For the Inviscid Conservative Gasdynamics Equations Using Flux-Vector Splitting", Report No. 80-4, August 1980, Flow Simulations, Inc.